Reducing InfoVis cluttering through non uniform sampling, displacement, and user perception

Enrico Bertini, Luigi Dell'Aquila, Giuseppe Santucci

Dipartimento di Informatica e Sistemistica - Università di Roma "La Sapienza"

Via Salaria, 113 - 00198 Roma, Italy - {bertini, dellaquila, santucci}@dis.uniroma1.it

ABSTRACT

Clutter affects almost any kind of visual technique and can obscure the structure present in the data even in small datasets, making it hard for users to find patterns and reveal relationships. In this paper we present a general strategy to analyze and reduce clutter using a special kind of sampling, together with an ad-hoc displacement technique and perceptual issues collected through a user study. The method, defined for 2D scatter plots, is flexible enough to be used in quite different contexts. In particular, in this paper we prove its usefulness against scatter plot, radviz, and parallel coordinates visualizations.

Keywords: visual clutter, quality metrics, data sampling, user study

1. INTRODUCTION

InfoVis techniques are often affected by clutter. The limits of visual devices, the adopted visualization, and the explored datasets, can easily generate cluttered images, especially when the number of items is high and/or the data present complex configurations.

Various solutions have been proposed to deal with such a problem; some of them try to resolve it locally, using tools that disambiguate the visualization in limited regions of the screen, e.g., fisheye lenses,[1] while others deal with it globally, that is, manipulating the whole representation in a way that permits to uncover potentially interesting features. In the latter category techniques such as: hierarchical clustering,[2] filtering,[3] constant density visualizations,[4] have proven to be useful, but they either require complex interactions or introduce major distortions that can affect the perception of interesting trends. Sampling has also been used as a "low cost" easy to implement technique and it forms the basis of our approach. But, simple uniform sampling suffers from some drawbacks that are discussed in the next section and addressed in our technique.

Here we follow the overall decluttering approach. In particular, our aim is to provide a "one shot" technique that permits, starting from a highly cluttered visualization, to automatically produce an uncluttered and accurate one. More specifically, we focus our attention on relative data densities, one of the most important clues a user looks at when analyzing a dataset's overall trend. Saturated areas that typically arise in cluttered visualizations, actually misrepresent density differences hiding potentially interesting patterns. Our aim is to make them apparent preserving, as much as possible, the existing dataset density differences.

Our proposal is based on the following main ideas:

- **Provide a formal framework -** we define a formal framework to measure the amount of degradation resulting from a given visualization, characterizing, e.g., the notion of density both in data and physical space. Using such a framework we define some quality metrics able to characterize the image degradation and to drive corrective actions.
- **Take into account density perception -** we analyze the way the users perceive density differences through a user study. This allows for fine tuning our metrics, considering perceptual density differences instead of numerical ones.

- **Provide a density-based decluttering algorithm** - we set up a set of strategies to reduce the image cluttering, using a specialized kind of sampling, i.e., non uniform sampling, that considers data density differences and applies different sampling rates to different image areas, trying to maximize the number of density differences available on the screen. The non uniform sampling strategy is enriched by a local pixel displacing strategy and incorporates perceptual issues.

Following this approach we set up a complete environment, proving its effectiveness in the context of 2D scatter plots. Encouraged by the results, we applied our method to other InfoVis techniques, namely radviz[5] and parallel coordinates.[6] The case of radviz is quite straightforward, while the parallel coordinates' one requires an additional adaptation effort that is described in Section 5. The results obtained in the three different cases let us believe that our strategy could be generalized to other techniques as well.

Summarizing, the contribution of the paper is twofold:

1. it presents a novel technique for reducing clutter that uses pixel displacements and non uniform sampling at the same time;

2. it exploits perceptual quality metrics for (a) estimating the image degradation and (b) driving automatic enhancing techniques.

The paper is structured as follows: Section 2 analyzes related works, Section 3 describes the user test results and introduces our model and quality metrics, Section 4 describes our sampling/displacing technique, Section 5 presents three case studies, and, finally, Section 6 presents conclusions and open problems.

2. BACKGROUND AND RELATED WORK

Our methods use quality metrics, user perception, sampling, and displacement, as a way to produce more accurate visualization and to reduce clutter. In the following we first report on related proposals on metrics for Information Visualization, then we relate our approach to clutter reduction to some other existing ones.

2.1. Metrics and perceptual issues

It is well known that Information Visualization needs metrics to provide precise indications on how effectively a visualization presents data. As expressed in[7] Information Visualization needs methods to measure the "*goodness*" of a given visualization and a definitive and strong set of methodologies/tools is still lacking.

First attempts towards this direction come from Tufte that in[8] proposes an interesting set of measures to estimate the "graphical integrity" of static (i.e., paper based) representations. Brath, in,[9] starting from Tufte's proposal, defines new metrics for static digital 3D images. He proposes metrics such as *data density* (number of data points/number of pixels) that recall Tufte's approach. He provides metrics aiming at measuring the visual image complexity like the *occlusion percentage*, that is the number of occluded elements in the visual space, or the *number of identifiable points*, that is the number of visible data points whose position is identifiable in relation to every other visible data point. While the main goal of the above metrics is to estimate a general visualization goodness, or to compare different visual systems, we mainly aim to assess the accuracy of a specific visualization, dealing with pixels and data points, in order to measure how accurately a visualization represents some data characteristics we are interested in (e.g., data density).

Measuring, in this context, means assessing the perception of visual features, therefore perceptual issues must be taken into account. Many studies have been conducted in the past to increase the effectiveness of visual systems and to avoid degradation. Results coming from color theory have been applied in practice in the context of data visualization to select color scales that appropriately reflect the properties of underlying data[10].[11] Preattentive features (visual features detected by the human eye without cognitive workload) have been exploited in the visualization of multivariate data to allow the users to efficiently detect visual patterns. Healey, in various proposals, inspected the effectiveness of preattentive features in depth and applied the results to build visually effective and efficient visualizations.[12] In this paper, we exploit the results of a perceptual study we conducted and presented in,[13] whose aim was to understand how users perceive density differences in point based 2D scatter plots. With these results, we are able to detect the threshold values beyond which density differences are perceived.

2.2. Dealing with clutter

The problem of reducing visual clutter to produce more effective visual representations has been directly and indirectly addressed by a variety of proposals. Some of them deal with the problem of reducing the overall clutter of the visualization, especially when the screen displays a large number of items, while others try to resolve clutter locally.

Jittering is used in commercial systems like Spotfire:[14] the overlapping items are displaced around their original position so that they become visible.[15] Trutschl et al. propose a smart jittering technique:[16] jittering is applied in a way that items that are similar in the n-dimensional data space are close to one another when moved from their original position. We also use a kind of jittering in our technique but, while the typical jittering displaces items around their original position randomly and applies it to the whole image, we use a selective jittering that runs only on specific areas and tries to move the items as little as possible from their original position.

Constant density visualization[4] is a distortion approach which is more oriented towards the representation of distorted overviews to deal with clutter. It presents more details within less dense areas and less details within denser ones, allowing the screen space to be optimally utilized and to reduce clutter. However, it is not clear if and how the method could scale to the case of large datasets.

Peng et al. in[17] propose dimension reordering as a complementary approach to reduce clutter in multivariate visualization. Computing the best order of axes according to some visual quality metrics, the visualization can make some patterns more evident without any information loss. Our approach is similar in that we also propose some visual quality metrics as driving parameter to run a decluttering algorithm, whereas the basis of our decluttering action is sampling.

In,[18] a density-based image processing technique is described. It permits to uncover clusters in parallel coordinates, mapping data density to color intensity. It has strong connections with our approach because we also use a density based approach. Moreover, the parallel coordinate case we describe in Section 5 draws from the above proposal the idea of analyzing parallel coordinates focusing on pairs of adjacent axes.

Sampling is also used as a way to reduce clutter. Since it reduces the number of displayed elements, the overall visual density decreases and the visualization becomes more intelligible. Uniform sampling has the interesting benefit that data features like distribution and correlation are preserved, allowing *"to see the overall trends in the visualization but at a reduced density"*.[19] Nonetheless, this idea is not free of drawbacks. In particular, choosing the right sampling rate is a challenging task and a straightforward solution does not exist, leaving to the user the burden of interactively selecting an appropriate sampling rate. Moreover, when data present both very high and very low density areas, two problems can arise:

- if sampling is too *strong* the areas in which density is very low become completely empty;
- if sampling is too *weak* the areas with highest densities still look all the same (i.e., completely saturated) and, consequently, density differences cannot be perceived.

In summary, our approach differs and relates to the proposals presented above in three main aspects: it provides a model for defining, both in a virtual and physical space, some metrics; it provides some *quantitative* information about an image quality; it exploits such results to automatically drive sampling algorithms preserving, as much as possible, specific visual characteristics.

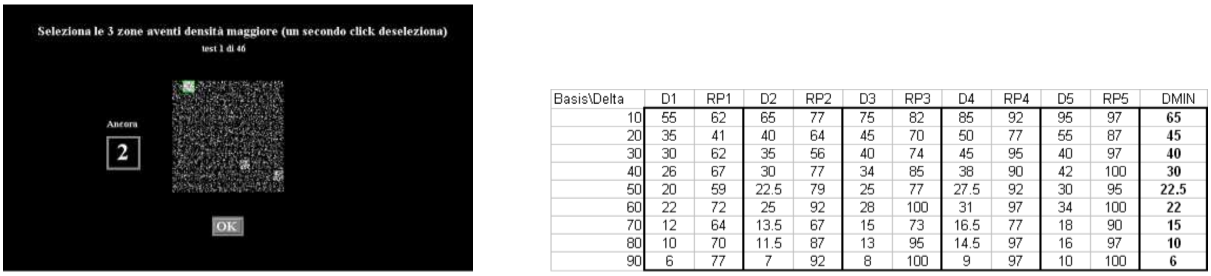
3. USER STUDY AND MODELING

In this section we present the results coming from the user study we performed in order to quantify the way a user perceives data densities in the context of 2D scatter plots. Moreover, we recall the main concepts of the formal framework we set up to characterize the notion of density and the metrics we use to estimate the image degradation.

3.1. User study

We performed a perceptual user study (details are in[13]) in order to answer the following question: what is the minimum difference in active pixels between two screen areas that allows a user to perceive a density difference?

The main idea underlying the experiment was to present the involved people with a uniform density screen (basis) containing three denser areas and to ask the users to recognize them. We repeated the test, for each subject, several times increasing both the uniform background density and the density difference (delta) between the background and the three denser areas. Figure 1 (a) shows a generic experiment step, in which the user marked one of the three denser areas.



Seleziona le 3 zone aventi densità maggiore (un secondo click deseleziona)

test 1 di 46

(a)

Basis\Delta	D1	RP1	D2	RP2	D3	RP3	D4	RP4	D5	RP5	DMIN
10	55	62	65	77	75	82	85	92	95	97	**65**
20	35	41	40	64	45	70	50	77	55	87	**45**
30	30	62	35	56	40	74	45	95	40	97	**40**
40	26	67	30	77	34	85	38	90	42	100	**30**
50	20	59	22.5	79	25	77	27.5	92	30	95	**22.5**
60	22	72	25	92	28	100	31	97	34	100	**22**
70	12	64	13.5	67	15	73	16.5	77	18	90	**15**
80	10	70	11.5	87	13	95	14.5	97	16	97	**10**
90	6	77	7	92	8	100	9	97	10	100	**6**

(b)

Figure 1. User study: (a)The user study's main screen (translation of the sentence in figure: "*Select the 3 areas with higher density (click twice to deselect)*"; (b) The user study results (all values are in percentages)

The results are collected in the table shown in Figure 1 (b). The table shows for each row a different basis (10, . . . , 90) and the five different incremental steps (δ) adopted in the test (D1, . . . , D5); for each increment the table shows the corresponding recognition percentage (RP1, . . . , RP5). As an example, the first row tells us that, while evaluating a basis of 10%, we asked the user to identify areas containing 55%, 65%, 75%, 85%, and 95% extra pixels and that the recognition rate was 62%, 77%, 82%, 92%, and 97%, respectively. We performed a single factor ANOVA test confirming the significance of our figures. The last column shows, for each basis, the minimum increment we have to choose to guarantee that 70 out of 100 end users will perceive the density difference. Using such a column we can interpolate a function $minimum\delta(A)$ returning the minimum density increment an area A' must show to be perceived as denser than A (by 70 out of 100 users).

The experiment results have been used to improve the accuracy of our algorithms, as described in the next section.

3.2. Models and metrics for density differences

We consider a 2D space in which we plot items by associating a pixel with each data element and the pixel position is computed mapping two data attributes on the spatial coordinates. Two data points are in collision when their projection is on the same physical pixel (in most cases that happens as a result of rounding) and free space is the percentage of pixels that is not assigned to any element. In[20] a framework to estimate the amount of colliding points and, as a consequence, the amount of free available space is described; here we report the main definitions useful for dealing with the algorithm discussed in Section 4 and the metrics described in the next subsection.

We assume the image is displayed on a rectangular area (measured in inches) and that small squares of area A divide the space in $m \times r$ *sample areas* (SA) where density is measured. Given a particular monitor, the resolution and size affect the values used in calculations. To fix the ideas, in the following examples we assume to use a monitor of 1280x1024 pixels and size of 13"x10.5". Using these figures we have 1,310,720 pixels available and if we choose SA of side $l = 0,08$ inch, the area is covered by 20.480 (160×128) sample areas whose dimension in pixels is 8×8.

For each $SA_{i,j}$, where $1 \leq i \leq m$ and $1 \leq j \leq r$, we calculate two different densities : *real data density* (data density in the following) and *represented density.*

Data density is defined as $d_{i,j} = \frac{n_{i,j}}{A}$ where $n_{i,j}$ is the number of data points that fall into sample area $A_{i,j}$. For a given visualization, the set of data densities is finite and discrete. In fact, if we plot n data elements, each $SA_{i,j}$ assumes a value $d_{i,j}$ within the set $0, \frac{1}{A}, \frac{2}{A}, \ldots, \frac{n}{A}$. In general, for any given visualization, a subset of these values will be really assumed by the sample areas. For each distinct value we can count the number of sample areas sharing the same data density. For example, if we plot 100 data points onto an area of 10 sample areas, we could have the following distribution: 3 sample areas with 20 data points, 2 sample areas with 15 data points, 2 sample areas with 5 data points.

Represented density is defined as $rd_{i,j} = \frac{p_{i,j}}{A}$ where $p_{i,j}$ is the number of distinct active pixels that fall into $SA_{i,j}$. The number of different values that a represented density can assume depends on the size of sample areas. If we adopt sample areas of 8×8 pixels the number of different non-empty densities represented is 64.

Because of collisions the number of active pixels on a sample area $SA_{i,j}$ will be smaller than the plotted point so $RD_{i,j} \leq D_{i,j} \; \forall i, j$.

3.2.1. Quality metrics

In the following we provide a quality metric that provides an indication of how many density differences are visible. Because of cluttering is associated with the number of collisions we focus on the distorted area disregarding the portion of the image that is correctly represented.

The complete list of the involved parameters is the following:

- the overall number of points being plotted, n;
- the image area size, in terms of number of pixels, x_pixels, y_pixels;
- the sample areas size in terms of number of pixels, l_pixels (we are considering squared sample areas);
- the number of collisions k per sample area (SA);
- the data density and the represented density.

Moreover, we introduce a threshold value Δ that allows for distinguishing the correctly represented part of the image from the distorted one. To fix the idea, we can state that we cannot bear SAs showing more than 32% of collisions w.r.t. $l\_pixels \times l\_pixels$.

Now we can concentrate on relative densities, measuring the preserved density differences through the metric PDDr (Preserved Data Densities ratio). This metric is calculated comparing pairs of sample areas and checking whether their relative data density (D) is preserved when considering their represented density (RD).

Introducing the $Diff(x, y)$ function defined as:

$$Diff(x, y) = \begin{cases} 1 & \text{if } x > y \\ 0 & \text{if } x = y \\ -1 & \text{if } x < y \end{cases}$$

we define the $match(i, j, k, l)$ function that returns true iff $Diff(D_{i,j}, D_{k,l}) = Diff(RD_{i,j}, RD_{k,l})$.

In order to produce a measure, we apply an algorithm that iteratively considers all pairs of Distorted SAs (DSA), comparing their D and RD through the $Diff$ function and counting the number of times it finds a matching pair.

Moreover, in order to take into account the relevance of a comparison between two sample areas, we weight each comparison using the number of points falling in the two sample areas.

In pseudo-code, the algorithm is:

```
function PDDr(){
 Let DSA[a][b]=distorted sample area in position (a, b)
 Let pairs=weighted pairs of distinct sample areas
 Let sum=weighted matching pairs of distinct sample areas

 pairs=0;
 sum=0;
 foreach distinct pair(DSA[i][j], DSA[k][l]){
   pairs = pairs + pt(DSA[i][j]) + pt(DSA[k][l]);
   if ( match(i, j, k, l) )
      sum = sum + pt(DSA[i][j])+ pt(DSA[k][l]);
 }
 return (sum / pairs);
}
```

where $pt(SA_{i,j})$ is a function returning the number of data points falling in a SA.

The variable *sum* contains the number of weighted matchings pairs encountered during the iterations; dividing it by the weighted total number of possible distinct distorted pairs we obtain the weighted percentage of matching sample areas ranging between 0 and 1 (the higher the better).

This metric provides a distortion evaluation, counting the densities differences still visible in the crowded area and weighting such differences through the involved points.

The main drawback of this metric is that it uses numerical differences between sample areas to decide whether a data density difference is well represented by the corresponding represented densities. As an example, a sample area containing 55 active pixels is considered denser than another one containing 54 active pixels while both of them look the same to the end user.

The experiment results allow to refine the metric, introducing the $PDiff(x, y)$ (Perceptual Diff) function as a modification of the above introduced $Diff(x, y)$ function:

$$PDiff(x, y) = \begin{cases} 1 & \text{if } x \geq y + y \times minimum\delta(y) \\ -1 & \text{if } y \geq x + x \times minimum\delta(x) \\ 0 & \text{otherwise} \end{cases}$$

Using the PDiff function within the match function, we obtain the PPDDr (*Perceptually* Preserved Data Densities ratio) metric. In this way the quality metric deals with user perceptible vs numeric density differences.

In order to better understand the difference between the two metrics, we apply them against the example in Figure 3(a) that shows about 160,000 mail parcels plotted on the X-Y plane according to their weight (X axis) and volume (Y axis). It is worth noting that, even if the screen is not fully saturated, the area close to the origin is very crowded (usually parcels are very light and small), so a great number of collisions is present in that area.

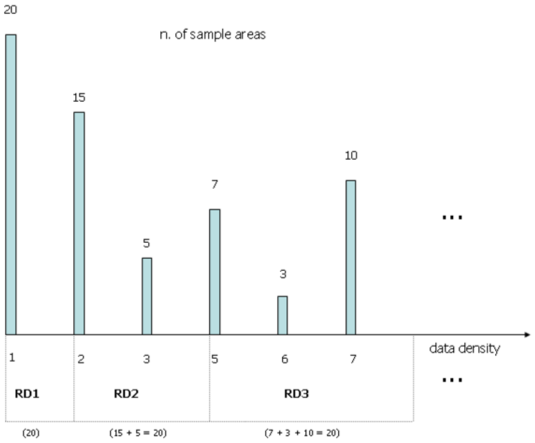
Plotting the data on a 600×600 pixels screen and using the pure numeric metric, PDDr with the 32% collision delta, we obtain the reasonable value of 0.71, meaning that in the distorted area about 71% of the data points are presented correctly in the image (i.e., their relative density is preserved in the final image). If we consider, by contrast, the PPDDr metric we obtain a worse (but more realistic) value, 0.57. That implies that the pure numeric metric counts a great number of "fake" density differences (numerical differences) not perceivable by the users.

4. CLUTTER REDUCTION

The problem we want to address is to represent, in the limited visualization space, as many density differences as we can, that is, trying to provide a visualization that is as representative as possible of the real underlying densities. In general, for visualizations where specific interventions are not employed, this correspondence is not accurate: when data are visualized on a crowded scatter plot, high data densities are mapped to few represented densities, the ones in which almost all pixels are active, and a large number of high data densities are "squeezed" on few and very closely represented densities; thus, some existing density differences cannot be perceived.

Any given visualization is a particular mapping between the set of data densities and the set of available represented densities, therefore the problem can be translated into the one of finding an optimal mapping between data densities and represented densities, that is, associating each data density to one of the represented densities. Modifying this mapping, we can potentially find some more accurate representations. To this aim, we must: (1) find a method to decide which data densities are mapped to which represented densities; (2) then we need a way to perform these mappings in practice.

As for the first point, we devised an algorithm that splits the range of existing data densities into 64 intervals which will be assigned to the 64 represented densities (under the hypothesis of 8×8 sample areas). To build such intervals the algorithm calculates the average number of non empty sample areas with respect to the represented densities ($K = \frac{n.sampleareas}{64}$). Starting from data density $D = 1$ ($D = 0$ is not taken into account in calculations because no interventions are applied in empty sample areas) it adds densities to the current interval until the sum of the included sample areas is equal to the average. When the value K is reached, a new interval is built. In the example in Figure 2 (a), $D = (1)$ is assigned to $RD = 1$ because it already spans 20 sample areas; $D = (2, 3)$ are assigned to $RD = 2$ because they sum up to 20; $D = (7, 3, 10)$ are assigned to $RD = 3$ because they sum up to 20; and so on. The resulting effect is that we have intervals of different sizes: large intervals containing many data densities that span few sample areas; short intervals containing few data densities that span many sample areas. The rationale behind this is that the algorithm tries to represent more accurately the densities that span a large portion of the screen while it accepts some distortion for few concentrated areas.



(a)

Represented Density	Perceptual Density
1	1
2, 3	2
4, 5, 6	4
7,8,9,10	7
11,12,13,14,15,16	11
17,18,19,20,21,22,23	17
24,25,26,27,28,29,30,31	24
32,33,34,35,36,37,38	32
39,40,41,42,43,44,45,46	39
47,48,49,50,51,52	47
53,54,55,56,57	53
58,59,60	58
61,62,63	61
64	64

(b)

Figure 2. Density split algorithm: (a)The data density axis is split into 64 values in a way that each one contains the same number of sample areas (b)Represented densities mapped to perceptual densities.

Once the mapping is computed we have, for each data density, a target density to reach. Thus, we need a way to turn on the number of pixels that produces the desired represented density. There are three possible cases and three associated interventions:

Represented density is equal to target density. This is the simplest case, the current represented density is already *equal* to the one we want to reach, so we just have to leave things as they are.

Represented density is greater than target density. This is the case in which the number of pixels turned on by the data points falling in the current sample area is *higher* than the target density we want to reach. In order to change the represented density we sample the data until we reach the target represented density.

Represented density is lower than target density. This is the case when the number of pixels turned on by the data points falling in the current sample area is *lower* than the target density we want to reach. It is worth to note that this case can happen only because of data points' collision. In order to reach the target density, we use pixel displacement so that some overlapping items become visible and represented density can be increased. In order to minimize the distortion introduced by displacement, the pixels are moved as close as possible to their original position. In any case, since the displacement takes place locally, within single sample areas, the entity of distortion is minimal and cannot have macroscopic negative effects.

When the right mapping has been performed and target densities are obtained we have, in principle, the best possible mapping. Looking at the image through the lens of our $PDDr$ metric (see Section 3) this is the best result we can achieve. But, as pointed out in Section 3, this trail of thoughts does not take into account the fact that little numerical differences cannot be perceived. This is why a third stage is needed.

Using the results given in Section 3, we restrict the values of represented densities to the ones that can be perceived as different. To this aim, we re-sample the dataset in a way that only the perceivable represented densities are presented on the screen, that is, the ones obtained in the user study. Starting from the represented density 64 downward to 1, they are sampled to let them reach their next available lower value: (1) to 1; $(2, 3)$ to 2; $(4, 5, 6)$ to 4; etc. The complete set of mappings is reported in Table 2 (b).

Roughly speaking, we can think of the whole process as follows. We have at disposal p different represented densities that are matched against k real data densities where, likely, $k >> p$; that implies that each represented density is in charge to represent several different data densities, hiding differences to the user. The strategy consists in changing, with sampling and displacement, the original data densities, altering their assignment to the p available represented densities to maximize the number of correctly represented data density differences. Then, since the problem of perceptual differences is recognized, an additional step is introduced in which the visualization is re-sampled to obtain only perceivable represented differences.

5. CASE STUDIES

In this section we test our approach against three InfoVis techniques, namely scatter plots, radviz, and parallel coordinates. It is worth noting that while in the first two cases the approach is quite straightforward, the parallel coordinates case requires a mapping between the two representations (presented in Section 5.3).

5.1. Scatter plot

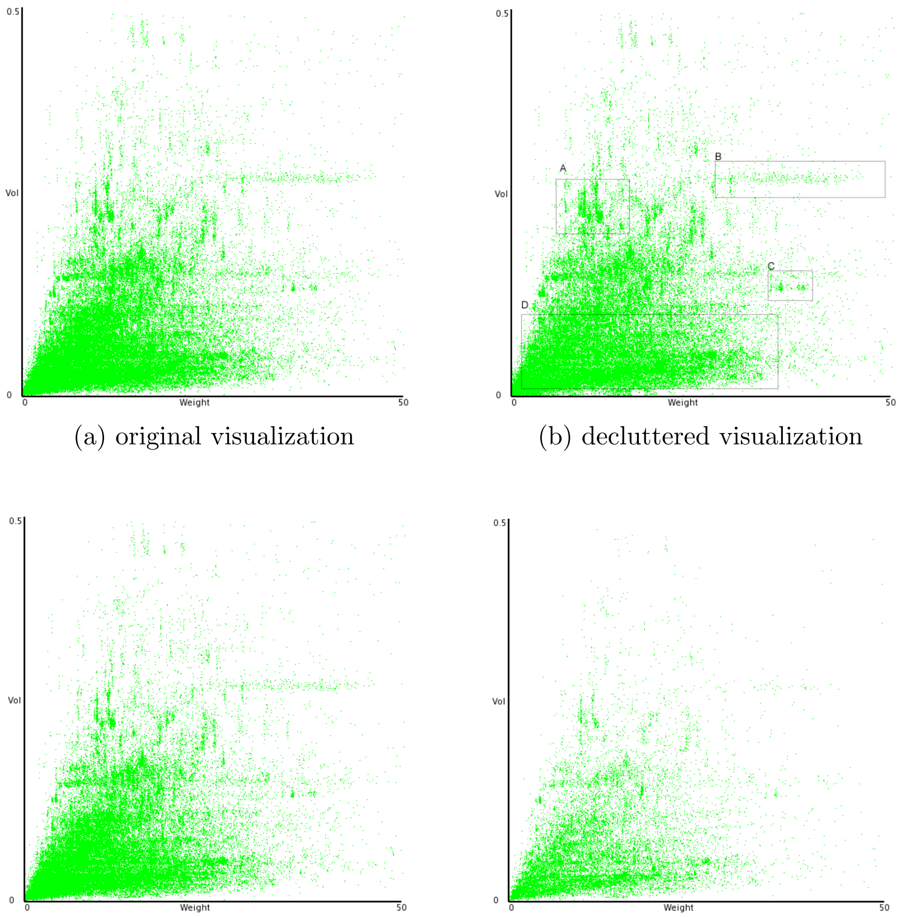
In order to assess the validity of our method we show the effect of its application to the parcel dataset in Figure 3. Figure 3 shows: (a) the original visualization, (b) the one obtained using our decluttering technique, (c) and (d) the visualization treated with plain uniform sampling applying high and low sampling factors, respectively.

The figure shows how the proposed method permits to quickly obtain an accurate and uncluttered representation: moving from the original visualization to the uncluttered one in one step. On the contrary, interactive uniform sampling would require the user to seek an appropriate sampling factor interactively.

Furthermore, the algorithm has the additional advantage of preserving patterns in low density areas and uncover interesting patterns in high density areas at the same time. This is a general characteristic of non uniform sampling; it can be better appreciated comparing the figures. Figure 3(d) shows the result of *strong* uniform sampling: some patterns become visible in the high density area in the lower left part of the image (D in the figure), but low density patterns like the one indicated as B in the figure get lost. Figure 3(c) shows the result of *soft* uniform sampling, where the opposite effect appears: low density patterns (D in the figure) are preserved but high density areas (B in the figure) still hide interesting patterns.

It is worth to note the effect of the displacement in terms of accuracy. Patterns A and C in Figure 3(b) reveal how the image can improve in terms of accuracy. While the small clusters in C remains almost unaltered in any visualization, the clusters in A becomes larger in Figure 3(b) because it hides a large number of overlapping items. Finally density differences more evident because not all the represented densities are available: according to Table 1, only 14 out of 64 represented densities are available on the screen and this, according to the user study, increases the density differences' perception.

The visual impressions are confirmed by our metrics: the original values of PDDr and PPDDr (0.71 and 0.57) increase in image (b) to 0.83 and 0.76. Our technique increased the percentage of data points that are correctly perceived by the end user (in terms of relative density) as much as 34% (i.e., 0.76/0.57).



(a) original visualization

(b) decluttered visualization

(c) *soft* uniform sampling

(d) *strong* uniform sampling

Figure 3. Comparison among sampling techniques: (a)original visualization; (b)visualization decluttered with our method (the boxes in figure refer to the discussion in Section 5.1); (c) and (d) *soft* and *strong* uniform sampling respectively.

5.2. Radviz

Radviz, from the point of view of visual perception, appears to be very similar to a scatter plot: each data point is represented through a single pixel on a 2D plane; the only difference is the particular rule of points' disposition. Concepts of data density, represented density, and clutter in a radviz representation are defined as for a scatter plot, so we can apply the same algorithm used for scatter plots to put in evidence differences of density in different areas of a radviz representation. To show the results obtained applying our technique on a radviz representation we introduce a five-dimensional dataset, named "out5d"; each data point represents an observation in a geographic region; dimensions are spot, magnetics, and three bands of radiometrics: potassium, thorium and uranium. Results of application of our technique are shown in Figure 4: in the original representation (on the left), differences of data density are not distinguishable. The sampled image (on the right) clearly shows some clusters and their differences of density (i.e., high density of central cluster and lower density of top-left horizontal cluster).

5.3. Parallel Coordinates

In order to challenge our approach and see if and how generalizable it is, we searched for a way to apply the same algorithm to a completely different visualization like parallel coordinates. Moving from the context of scatter plots to the one of parallel coordinates is not an easy task. It requires moving from a visualization where each data item is represented by a single pixel, to one where each data item is represented by a polyline. Moreover,

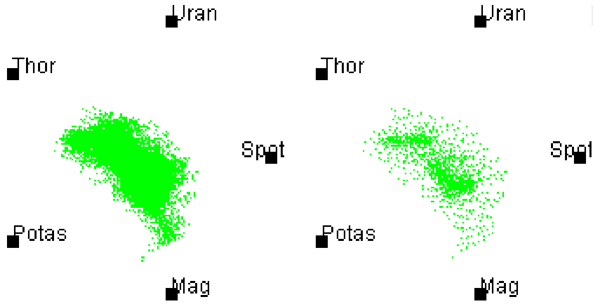


Figure 4. Applying sampling to radviz (Out5d dataset). Original radviz representation (left) and sampled data representation (right). Sampling extracts clusters and puts in evidence differences of density between different ones

while scatter plots allow only for two geometric attributes, parallel coordinates permits, in principle, to show as many geometric attributes as one wants.

One possibility is to redesign the concepts developed above, e.g., data density, represented density, sample areas, etc., in the new context. While this is a viable and potentially interesting way, we favored a different approach. Inspired by the mathematical concept of "transform" (e.g., fourier transform) we decided to explore the idea of mapping the new visualization (parallel coordinates in our case) to the one we already know how to handle, that is, scatter plots, apply the decluttering method, and step back to the original display to see the results. Note that the idea has reasonable foundations as long as: (1) sampling is applied on data items, therefore the elimination of one data item corresponds to the elimination of one single mark in any visualization; (2) the displacement of an item in the transformed visualization produces a displacement in the original one. The mapping between parallel coordinates and scatter plots requires two choices: one is how to project a multidimensional visualization on a bi-dimensional one, another is how to choose the resolution of the scatter plot.

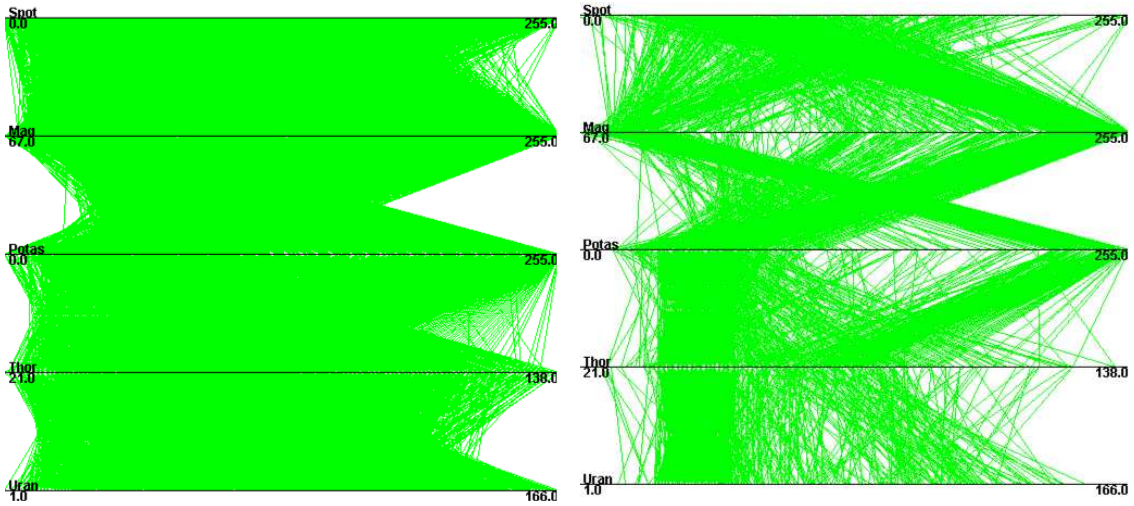
As for the first choice, we adopted the approach of mapping every single couple of adjacent parallel coordinates' axes, to a single scatter plot, so that for a parallel coordinate visualization with, e.g., three axes we have two scatter plots. This implies, in general, that different pairs will be treated with different runs of the algorithm. Treating pairs of axes in isolation has been effectively adopted in other proposals, like in,[18] and it has proven to be effective in indicating interesting data trends.

For what regards the selection of resolution, we adopted the straightforward approach of using a scatter plot whose axes cover the same number of pixels of the parallel coordinates' axes, that is, if the couple of axes $X1$, $X2$ have a pixel height of P pixels, the resulting scatter plot will use an area of $P \times P$ pixels.

This mapping is quite straightforward to implement and preserves the notion of collision and displacement: a displacement in the scatter plot will produce a displacement of the corresponding lines between the two axes. Two overlapping pixels in the scatter plot, represent two perfectly overlapping segments between the two axes in the parallel coordinates.

Summarizing, we proceed as follows:

- starting from the n-dimensional dataset we create n-1 bi-dimensional datasets projecting the data on the n-1 pairs of *adjacent* parallel coordinates;
- we represent each new dataset DS_i on a $P \times P$ pixels scatter plot;
- we apply the sampling/displacement algorithm to each scatter plot (disregarding the perceptual issues);
- we represent in parallel coordinates only the segments corresponding to pixels belonging to the sampled scatter plots.



(a) original

(b) decluttered

Figure 5. Parallel coordinates case: (a) the original visualization; (b) visualization obtained with our mapping method.

Concerning the sampling algorithm, since the n-1 scatter plots are just an internal data representation, we do not consider the perceptual issues and we allow the algorithm to use all the available represented densities.

Applying the above procedure we obtain different samplings on different pairs of dimensions; the corresponding partial drawing of the polylines allows for discovering general relationships among attributes. As an example, in Figure 5 we can see (a) the original "out5d" visualization and (b) the result of mapping/sampling activities described so far. Some patterns among data are now quite evident; e.g., high Magnesium values correspond to low Potassium values and viceversa. Moreover high Magnesium values present low values for Thorium and Uranium as well.

6. CONCLUSION AND FUTURE WORK

In this paper we presented a novel automatic strategy for enhancing InfoVis images, preserving in the final visualization as many density differences as possible. To this aim, the strategy incorporates three novel techniques:

1. it uses sampling and pixel displacement at the same time;
2. it is driven by perceptual issues gathered from a suitable user study;
3. it incorporates a quality metric useful for both driving the algorithm and validating the results.

The technique effectiveness has been proved applying it to three InfoVis techniques, 2D scatter plot, radviz, and parallel coordinates.

Several open issues rise from this work:

- the mapping between parallel coordinates and scatter plot deserves a deeper analysis and a fine tuning;
- both the notions of collision and density should be refined in the context of parallel coordinates;
- user perception of density in the context of parallel coordinate should be investigated.

We are currently working on these issues and we are trying to generalize the mapping approach followed for the parallel coordinates in order to devise some general guidelines for applying our strategy to other, more complex, InfoVis techniques.

REFERENCES

1. M. S. T. Carpendale and C. Montagnese, "A framework for unifying presentation space," in *Proc. of UIST Symposium on User interface Software and Technology*, pp. 61–70, ACM Press, (New York, NY, USA), 2001.
2. Y.-H. Fua, M. O. Ward, and E. A. Rundensteiner, "Navigating hierarchies with structure-based brushes," in *Proc. of the 1999 IEEE Symposium on Information Visualization*, p. 58, IEEE Computer Society, 1999.
3. C. Ahlberg and B. Shneiderman, "Visual information seeking: tight coupling of dynamic query filters with starfield displays," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 313–317, 1994.
4. A. Woodruff, J. Landay, and M. Stonebraker, "Constant density visualizations of non-uniform distributions of data," in *Proceedings of the 11th annual ACM symposium on User interface software and technology*, pp. 19–28, ACM Press, 1998.
5. P. Hoffman, G. Grinstein, and D. Pinkney, "Dimensional anchors: a graphic primitive for multidimensional multivariate information visualizations," in *NPIVM '99: Proc. Workshop on New Paradigms in Information Visualization and Manipulation*, pp. 9–16, ACM Press, 1999.
6. A. Inselberg and B. Dimsdale, "Parallel coordinates: a tool for visualizing multi-dimensional geometry," in *VIS '90: Proceedings of the 1st conference on Visualization '90*, pp. 361–378, IEEE Computer Society Press, 1990.
7. N. Miller, B. Hetzler, G. Nakamura, and P. Whitney, "The need for metrics in visual information analysis," in *Proceedings of the 1997 workshop on New paradigms in information visualization and manipulation*, pp. 24–28, ACM Press, 1997.
8. E. R. Tufte, *The visual display of quantitative information*, Graphics Press, 1986.
9. B. Richard, "Concept demonstration: Metrics for effective information visualization," in *Proceedings For IEEE Symposium On Information Visualization*, pp. 108–111, IEEE Service Center, Phoenix, AZ, 1997.
10. H. Levkowitz and G. T. Herman, "Color scales for image data," *IEEE Computer Graphics and Applications* **12**(1), pp. 72–80, 1992.
11. L. D. Bergman, B. E. Rogowitz, and L. A. Treinish, "A rule-based tool for assisting colormap selection," in *Proceedings of the 6th conference on Visualization '95*, p. 118, IEEE Computer Society, 1995.
12. C. G. Healey and J. T. Enns, "Large datasets at a glance: Combining textures and colors in scientific visualization," *IEEE Transactions on Visualization and Computer Graphics* **5**(2), pp. 145–167, 1999.
13. E. Bertini and G. Santucci, "Is it darker? improving density representation in 2d scatter plots through a user study," in *Proc. of SPIE Conference On Visualization and Data Analysis*, January 2005.
14. C. Ahlberg and E. Wistrand, "Ivee: an information visualization and exploration environment," in *Proc. of IEEE Symposium on Information Visualization*, p. 66, IEEE Computer Society, 1995.
15. J. Manson, "Occlusion in two-dimensional displays: Visualization of meta-data," tech. rep., University of Maryland, College Park, 1999.
16. M. Trutschl, G. Grinstein, and U. Cvek, "Intelligently resolving point occlusion," in *Proceedings of the IEEE Symposium on Information Vizualization 2003*, p. 17, IEEE Computer Society, 2003.
17. W. Peng, M. O. Ward, and E. A. Rundensteiner, "Clutter reduction in multi-dimensional data visualization using dimension reordering," in *Proc. of IEEE Symposium on Information Visualization*, pp. 89–96, IEEE Computer Society, 2004.
18. A. O. Artero, M. C. F. de Oliveira, and H. Levkowitz, "Uncovering clusters in crowded parallel coordinates visualizations," in *Proc. of IEEE Symposium on Information Visualization*, pp. 81–88, IEEE Computer Society, 2004.
19. G. Ellis and A. Dix, "by chance: enhancing interaction with large data sets through statistical sampling.," in *Proc. of ACM Working Conference on Advanced Visual Interfaces*, pp. 167–176, ACM Press, may 2002.
20. E. Bertini and G. Santucci, "By chance is not enough: preserving relative density through non uniform sampling," in *Proc. of IEEE Information Visualization Conference*, July 2004.